PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2003-177781**
(43)Date of publication of application : **27.06.2003**

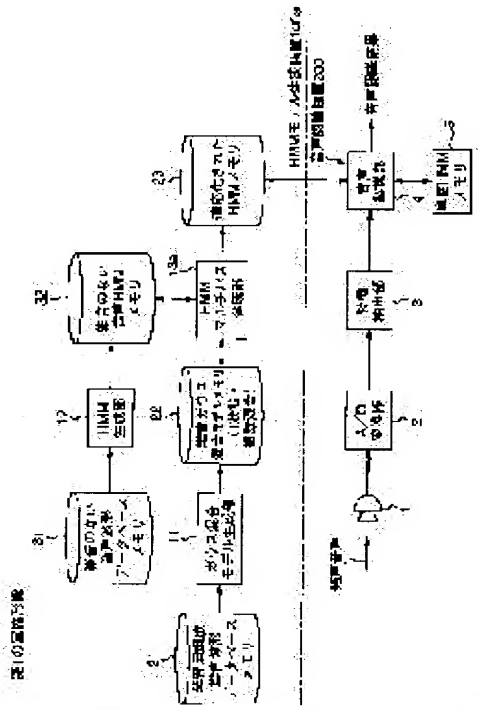
(51)Int.Cl. G10L 15/06
G10L 15/10
G10L 15/14

(21)Application number : **2001-378546**
(22)Date of filing : **12.12.2001**

(71)Applicant : **ADVANCED TELECOMMUNICATION RESEARCH INSTITUTE INTERNATIONAL**
(72)Inventor : **IDA MASAKI**
NAKAMURA SATORU

(54) ACOUSTIC MODEL GENERATOR AND VOICE RECOGNITION DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To provide an acoustic model which is not restricted by a condition that the SN ratio of an input voice is known.
SOLUTION: A Gaussian mixture model generation part 11 generates a multiple mixed Gaussian mixture model in one state on the basis of the waveform signal data of a plurality of kinds of environmental noise for learning which are stored in a database 21, so that an output likelihood can be maximum, and an HMM synthesizing part 13 generates a plurality of adapted HMMs which include mixture Gaussian distributions in respective states, which are represented by the sum of linear coupling of Gaussian distributions weighted by a prescribed weight coefficient in all combined states of respective states and correspond to a plurality of SNs between the noiseless voice HMM and the generated noise Gaussian mixture models, in accordance with a prescribed noiseless voice HMM and the generated noise Gaussian mixture models, and the plurality of generated adapted HMMs are juxtaposed to generate an acoustic model in a multipath form. A voice recognition part 4 uses the adapted acoustic model to perform voice recognition of an uttered voice signal on the basis of an extracted feature quantity.



LEGAL STATUS

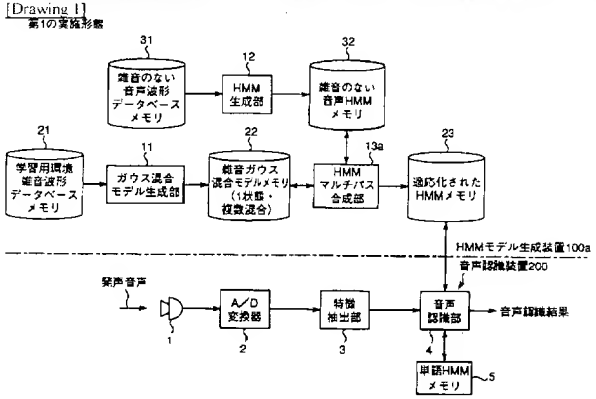
[Date of request for examination] 07.10.2003
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

* NOTICES *

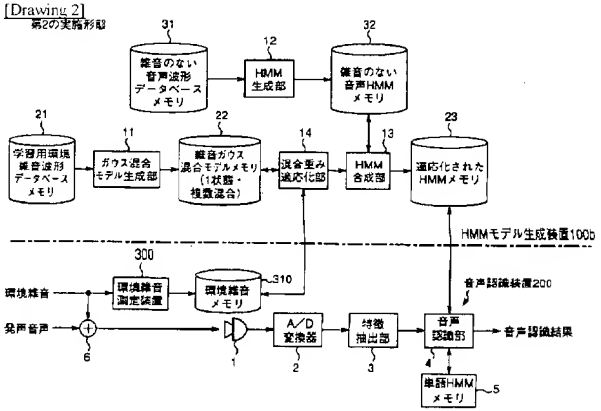
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

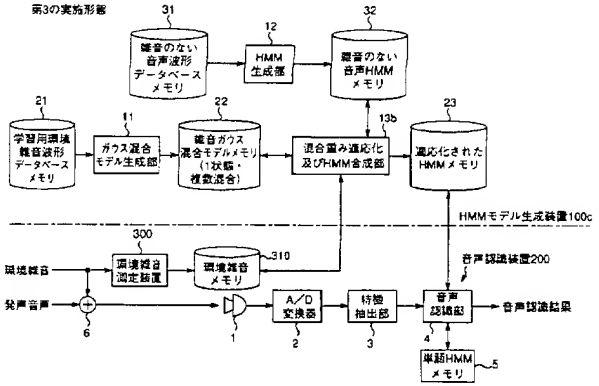
[Drawing 1]



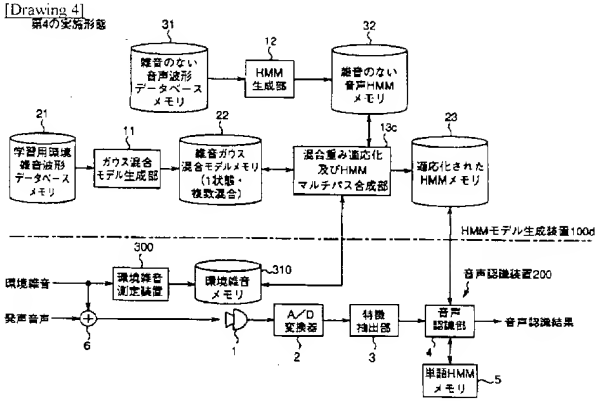
[Drawing 2]



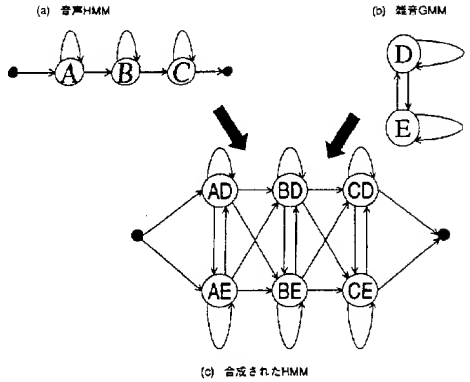
[Drawing 3]



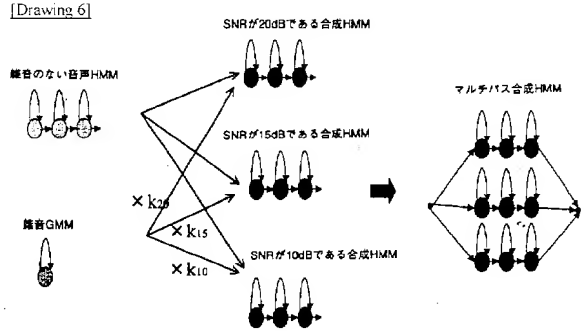
[Drawing 4]



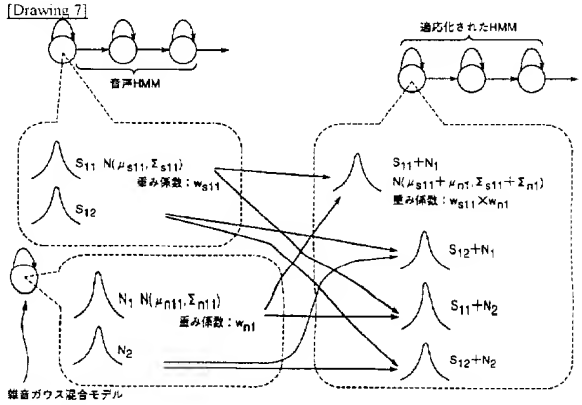
[Drawing 5]



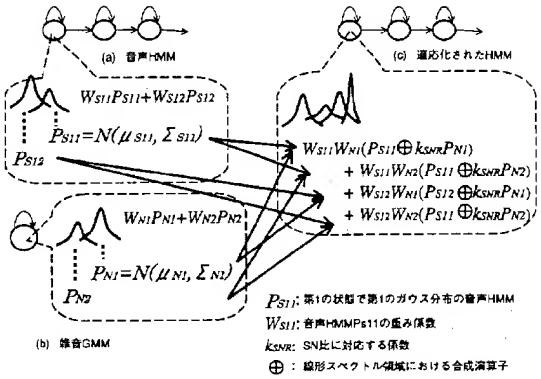
[Drawing 6]



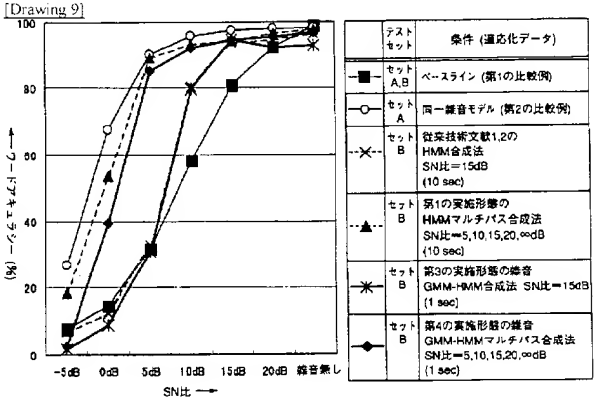
[Drawing 7]



[Drawing 8]



[Drawing 9]



[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram showing the configuration of HMM model generation equipment 100a which is the 1st operation gestalt concerning this invention, and a voice recognition unit 200.
[Drawing 2] It is the block diagram showing the configuration of HMM model generation equipment 100b which is the 2nd operation gestalt concerning this invention, a voice recognition unit 200, and the ambient noise measuring device 300.
[Drawing 3] It is the block diagram showing the configuration of HMM model generation equipment 100c which is the 3rd operation gestalt concerning this invention, a voice recognition unit 200, and the ambient noise measuring device 300.
[Drawing 4] It is the block diagram showing the configuration of 100d of HMM model generation equipment which is the 4th operation gestalt concerning this invention, a voice recognition unit 200, and the ambient noise measuring device 300.
[Drawing 5] It is the explanatory view showing how to compound Voice HMM and Noise GMM by the HMM synthesis method concerning the 3rd conventional example.
[Drawing 6] It is the explanatory view by the HMM multi-pass synthesis method concerning the 1st operation gestalt showing the voice HMM without a noise, and the approach of compounding Noise GMM.
[Drawing 7] It is the explanatory view showing derivation of the output probability distribution after HMM composition when Voice HMM and a noise gauss mixing model are expressed by output distribution of 2 mixing.
[Drawing 8] It is the explanatory view showing the approach of the formation of mixed weight adaptation used with the 2nd thru/or 4th operation gestalt, and HMM composition.
[Drawing 9] It is an experimental result concerning the 1st example of a comparison, the 2nd example of a comparison, the conventional technical reference 1 and 2, the 1st operation gestalt, the 3rd operation gestalt, and the 4th operation gestalt, and is the graph which shows the WORD accuracy to an SN ratio.
[Description of Notations]
1 -- Microphone,
2 -- A/D converter
3 -- Feature-extraction section,
4 -- Speech recognition section,
5 -- Word HMM memory,
6 -- Adder,
11 -- Gauss mixing model generation section,
12 -- HMM generation section,
13 -- HMM composition section,
13 a--HMM multi-pass composition section,
13b -- The formation of mixed weight adaptation, and the HMM composition section,
13c -- The formation of mixed weight adaptation, and the HMM multi-pass composition section,
14 -- Mixed weight adaptation-ized section,
21 -- Ambient noise data-point Base Memory for study,
22 -- Gauss mixing model memory,
HMM memory formed into 23 -- adaptation,
31 -- Voice data-point Base Memory without a noise,
32 -- Voice HMM memory without a noise,
100a, 100b, 100c, 100 d--HMM model generation equipment,
200 -- Voice recognition unit,
300 -- Ambient noise measuring device,
310 -- Ambient noise memory.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the sound model generation equipment and the voice recognition unit for a voice recognition unit.
[0002]
[Description of the Prior Art] When the use under the real environment of a voice recognition system is considered, since a surrounding environmental sound exists, recognition performance degradation is not avoided. Then, a robust sound model is needed to mixing of a surrounding environmental sound. Since the environmental sound at the time of recognition itself cannot be used for mixing of an environmental sound as a method of generating a robust sound model, the method of predicting a mixing environmental sound beforehand and performing adaptation-ization is used. However, since prediction of a mixing environmental sound contains a fluctuation component, there are many difficult things.
[0003] It is divided roughly into the following two as the approach of adaptation-izing of the sound model in the conventional technique. One side is the approach of creating the sound model which assumed the environmental sound at the time of recognition at the time of a system design. That is, for example, after generating the noise hidden Markov model for study (a hidden Markov model is hereafter called HMM.) based on the wave database of a known noise, using this, by learning the voice HMM without a noise, adaptation-ized HMM can be generated and this can be used for a voice recognition unit (henceforth the 1st conventional example).
[0004] On the other hand, another side is a method for which a sound model is adapted at any time with the environmental sound data at the time of recognition. Since the environmental sound at the time of recognition itself cannot be used for adaptation-ization, generally the comparatively little environmental sound in front of voice input is used (henceforth the 2nd conventional example).
[0005] According to the approach of the 1st conventional example, strong robustness is shown to mixing of the environmental sound within the assumed limits. However, since it is necessary to take into consideration about the combination of all voice and environmental sounds when it cannot respond to a strange noise, but there is a trouble that robustness is missing and mixing of various environmental sounds is assumed, in a cost side, it is not realistic. That is, when the class of known noise was made [many], there was a trouble that the computational complexity of adaptation-ized HMM became great.
[0006] In the approach of the 2nd conventional example, it is very difficult to predict all the environmental sounds under recognition from little data, and cannot respond to mixing of the environmental sound besides an assumption.
[0007] The constraint that the description of an environmental sound that all the environmental sounds that mix the former mix the conditions of being known, and the latter is eternal exists. Generally, since the environmental sound contains the component to change in real use, the above-mentioned constraint is not necessarily fulfilled.
[0008] In order to solve the above trouble, this invention persons In the patent application of an application for patent No. 283516 2000 to], based on the wave signal data of the ambient noise of two or more classes for "study, so that output likelihood may serve as max Generate the gauss mixing model of two or more mixing in the one condition, and the noise gauss mixing model (GMM) to which generation was carried out [voice / HMM / without a predetermined noise / above-mentioned] is set in the condition of all the combination of each of these conditions. Sound model generation equipment compounded by generating adaptation-ized HMM including the mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor" (it is hereafter called the 3rd conventional example.) It has proposed.
[0009] The HMM synthesis method used in this 3rd conventional example is compounding the sound model of the phoneme which learned using the voice which does not have a noise in advance, and the model of ambient noise, and is the approach of creating the sound model which was adapted for the modeled ambient noise. Here, only the noise of additivity is assumed. The power spectrum of the input voice observed is set to Y, and the power spectrum N of ambient noise and the power spectrum S of clean voice without a noise express this, and the amount of the linearity spectral region in each parameter is given to those notations, and it expresses with the bottom "linspc" here. The additivity of ambient noise is materialized like a degree type in a linearity spectral region.
[0010]
[Equation 1]
Ylinspc=Slinspc+Nlinspc [0011] On the other hand, since the feature extraction of the sound model is generally carried out with the spectrum, it serves as a degree type.
[0012]
[Equation 2] Ycep=gamma -1, log [exp{gamma(Scep)}+k-exp {gamma (Ncep)}]
[0013] gamma is the operator of the Fourier transform, gamma-1 is the operator of inverse Fourier transform here, and k is a multiplier determined according to a signal-to-noise power ratio (henceforth an SN ratio). It is expressed with each direct product of HMM as the structure of Composition HMM shows two above to drawing 5 , when adapted for HMM. Transition probability is searched for by the corresponding product of transition probability, and output probability distribution is combined in each condition.

[0014]
[Problem(s) to be Solved by the Invention] However, even if it was the case where speech recognition was carried out using the sound model generated with the sound model generation equipment concerning the 3rd conventional example, the rate of speech recognition still had the trouble of being low.
[0015] Moreover, in the HMM synthesis method used in the 3rd conventional example, there is constraint that the SN ratio of input voice is known as shown in two above. If this constraint is removable, about an SN ratio, a degree of freedom is high and can generate the sound model which can respond to various SN ratios.
[0016] It is in the purpose of this invention offering the sound model generation equipment which generates the sound model which can obtain the high rate of speech recognition as compared with the 3rd conventional example, and the voice recognition unit using the sound model generation equipment concerned, without solving many above troubles, being dogged and increasing the computational complexity of a sound model to mixing of a strange noise.
[0017] Moreover, it is in the purpose of this invention offering the sound model generation equipment which generates the sound model which is not restrained by constraint that the above trouble is solved and the SN ratio of input voice is known, and the voice recognition unit using the sound model generation equipment concerned.
[0018]
[Means for Solving the Problem] The sound model generation equipment concerning invention of the 1st of this application From the wave signal data of the ambient noise of two or more classes for study stored in a storage means to store the wave signal data of the ambient noise of two or more classes for study, and the above-mentioned storage means, so that output likelihood may serve as max In the condition of all the combination of a generation means to generate the gauss mixing model of two or more mixing in the one condition, the voice HMM without a predetermined noise, and the noise gauss mixing model generated by the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. And two or more adaptation-ized HMM(s) corresponding to two or more signal-to-noise ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / without the above-mentioned noise / above-mentioned] are generated. It is characterized by having a synthetic means to generate the sound model which compounds and becomes so that adaptation-ized HMM of the generated plurality may be juxtaposed and it may become a multi-pass format.
[0019] Moreover, the sound model generation equipment concerning invention of the 2nd of this application From the wave signal data of the ambient noise of two or more classes for study stored in a storage means to store the wave signal data of the ambient noise of two or more classes for study, and the above-mentioned storage means, so that output likelihood may serve as max In the condition of all the combination of a generation means to generate the gauss mixing model of two or more mixing in the one condition, the voice HMM without a predetermined noise, and the noise gauss mixing model generated by the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. And it is characterized by having a synthetic means to generate the sound model which becomes in HMM adaptation-ized in the mixed weight of the above-mentioned noise gauss mixing model based on the ambient noise data at the time of speech recognition.
[0020] Furthermore, the sound model generation equipment concerning invention of the 3rd of this application From the wave signal data of the ambient noise of two or more classes for study stored in a storage means to store the wave signal data of the ambient noise of two or more classes for study, and the above-mentioned storage means, so that output likelihood may serve as max In the condition of all the combination of a generation means to generate the gauss mixing model of two or more mixing in the one condition, the voice HMM without a predetermined noise, and the noise gauss mixing model generated by the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. Based on the ambient noise data at the time of speech recognition, mixed weight of the above-mentioned noise gauss mixing model is adaptation-ized. And two or more adaptation-ized HMM(s) corresponding to two or more signal-to-noise ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / without the above-mentioned noise / above-mentioned] are generated. It is characterized by having a synthetic means to generate the sound model which compounds and becomes so that adaptation-ized HMM of the generated plurality may be juxtaposed and it may become a multi-pass format.
[0021] Furthermore, the voice recognition unit concerning invention of the 4th of this application is characterized by to have an extract means extract the characteristic quantity based on the utterance sound signal of a natural utterance sentence, and a speech-recognition means performs speech recognition of the above-mentioned utterance sound signal using the sound model generated by claim 1 thru/or the sound model generation equipment of one of any 3 publications based on the characteristic quantity by which the extract was carried out [above-mentioned], and output a speech-recognition result.
[0022]
[Embodiment of the Invention] Hereafter, the operation gestalt which starts this invention with reference to a drawing is explained.
[0023] <Operation gestalt of ** 1st> drawing 1 is the block diagram showing the configuration of HMM model generation equipment 100a which is the 1st operation gestalt concerning this invention, and a voice recognition unit 200. HMM model generation equipment 100a concerning this operation gestalt The gauss mixing model generation section 11 which generates a noise gauss mixing model using the ambient noise wave database for study including the noise wave of the environmental sound of two or more classes, The voice HMM without a noise using the generated noise gauss mixing model Two or more adaptation-ized HMM(s) corresponding to two or more SN ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / which learns with a well-known HMM synthesis method, and does not have the above-mentioned noise / above-mentioned] are generated. It has HMM multi-pass composition section 13a which generates the sound model of HMM which compounds and becomes so that adaptation-ized HMM of the generated plurality may be juxtaposed and it may become a multi-pass format as a main component. In order to raise robustness when it is the approach of building a robust voice model by small computational complexity and a strange environmental sound specifically mixes in an environmental variation with this operation gestalt, It assumes that various environmental sounds mix beforehand, and

environmental adaptation-ization by the HMM composition which gave the environmental sound of two or more classes as adaptation data is performed. Here An environmental sound is independently learned as HMM. A well-known HMM synthesis method (For example) "The conventional technical reference 1 F.Martin et al., "Recognition of Noisy Speech by Composition of Hidden Markov Models", an Institute of Electronics, Information and Communication Engineers technical report, SP 92-96, pp.9-16, 1992", reference, such as the conventional technical reference 2 "the "likelihood maximization adapted method based on HMM composition" besides South Yasuhiro, an Institute of Electronics, Information and Communication Engineers technical report, and SP95-June, 1995 [24 or]." All voice models are made to adaptation-ize effect of the environmental sound of two or more classes.

[0024] The technique used with this operation gestalt in order to solve the above-mentioned trouble in the conventional technique performs environmental adaptation-ization assumed that various environmental sounds mix beforehand in order to raise robustness when a strange environmental sound mixes. Various environmental sounds are independently learned as a noise gauss mixing model, and it becomes possible by making all voice models adaptation-ize effect of the environmental sound of two or more classes by HMM composition to build a robust voice model by small computational complexity to an environmental variation.

[0025] Moreover, in the HMM synthesis method in the 3rd conventional example, there is constraint that the SN ratio of input voice is known as shown in two above. The technique of building the adaptation-ization HMM corresponding to two or more SN ratios to juxtaposition is used for solution of this problem. The explanatory view of this technique is shown in drawing 6 . By this technique, in case a noise model is compounded, two or more composition HMM (multi-pass model classified by SN ratio) corresponding to some SN ratios within the limits predicted as input voice is obtained (with this operation gestalt, it is SN ratio =10 and 15 or 20dB, and this technique is called below "HMM multi-pass synthesis method".). Since the SN ratio of input voice is not known in the case of speech recognition, each [these] composition HMM is dealt with as one model. That is, in case the pass of two or more SN ratios is defined and decoded to each model, it constitutes so that the path in which likelihood is the highest may be made to choose.

[0026] In drawing 1 , the database of two or more speakers' large-scale voice wave signal with a phoneme label (it is [no noise] and is clean.) is stored, and the HMM generation section 12 outputs and stores voice data-point Base Memory 31 without a noise in the voice HMM memory 32 which generates the voice HMM without a noise and does not have a noise so that output likelihood may serve as max using well-known EM (Expectation-Maximization) algorithm based on the database concerned. On the other hand, ambient noise data-point Base Memory 21 for study For example, the Japan Electronic Industry Development Association noise database (for example) reference, such as the conventional technical reference 3 "the Japan Electronic Industry Development Association noise database, Japan Electronic Industry Development Association, and http://www.jeida.or.jp/committee/humanmed/speech/noisedbj.html." The data of the wave signal of the ambient noise of two or more stored classes for study are stored. It is based on the data of the wave signal of the ambient noise of two or more classes for study stored in this database memory 21. The gauss mixing model generation section 11 Using well-known EM algorithm, the noise gauss mixing model of two or more mixing is generated in the one condition, and it outputs and stores in the noise gauss mixing model memory 22 so that output likelihood may serve as max. Furthermore, the voice HMM in which HMM multi-pass composition section 13a does not have the noise stored in the voice HMM memory 32, A well-known HMM synthesis method is used from the noise gauss mixing model stored in the model memory 22. Are adaptation-ized HMM and two or more adaptation-ized HMM(s) corresponding to two or more signal-to-noise ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / without the above-mentioned noise / above-mentioned] are generated. It outputs and stores in the HMM memory 23 which compounds and becomes so that adaptation-ized HMM of the generated plurality may be juxtaposed and it may become a multi-pass format and which generated adaptation-ized HMM and was adaptation-ized.

[0027] The HMM synthesis method used by HMM multi-pass composition section 13a of this operation gestalt is the approach of compounding the voice HMM learned in the clean environment where a noise does not exist, and the noise gauss mixing model which learned the description of an environmental sound, and creating HMM to the voice which the environmental sound mixed. carrying out cosine conversion of each Gaussian distribution of voice and a noise in a cepstrum field in this HMM synthesis method, respectively so that it may be illustrated by drawing 2 of the conventional technical reference 2 -- a logarithm -- carrying out characteristic conversion further, after changing into each Gaussian distribution of the voice of a spectral region, and a noise -- the voice of a linearity spectral region, and the logarithm of a noise -- it changes into Gaussian distribution. here -- the voice of the linearity spectral region after characteristic conversion, and the logarithm of a noise -- the logarithm of the voice which the noise in a linearity spectral region superimposed by carrying out weighting-factor attachment addition of the Gaussian distribution mutually -- Gaussian distribution is generated. furthermore, the logarithm of the voice which the generated noise superimposed -- Gaussian distribution -- logarithmic transformation -- carrying out -- a logarithm -- after changing into the Gaussian distribution of the voice which the noise in a spectral region superimposed, the Gaussian distribution of the voice which the noise in a cepstrum field superimposed is acquired by carrying out inverse cosine conversion further. The above is the synthesis method of the output probability in a HMM synthesis method.

[0028] The condition of a noise gauss mixing model expresses output probability distribution with mixed Gaussian distribution, in order to express the versatility of an environmental sound. The output probability distribution of HMM after the composition at this time is expressed with the sum of the mixed distribution of Voice HMM, and the mixed distribution of a noise gauss mixing model in a cepstrum field. That is, each Gaussian distribution which constitutes mixed distribution is expressed by the sum of all the combination in the Gaussian distribution of each condition of Voice HMM, and the condition of a noise gauss mixing model, and a mixed weighting factor is expressed by the product of each weighting factor.

[0029] As mentioned above, when performing composition with Voice HMM and a noise gauss mixing model and each output distribution is expressed by mixed Gaussian distribution, the output distribution after composition becomes all the combination of each mixed element. The average and distribution of each element after composition become the sum of the original mixed element. The mixed weighting factor of each element after composition is expressed with the product of the original mixed weighting factor. Drawing 7 shows derivation of the output probability distribution after HMM composition when Voice HMM and a noise gauss mixing model are expressed by output distribution of 2 mixing. In addition, in drawing 7 , N (-) shows an average and distribution of each Gaussian distribution. The output probability distribution of the 1st condition of Voice HMM is the sum with a weighting factor of Gaussian distribution S11 and S12, and the output probability distribution of a noise gauss mixing model is the sum with a weighting factor of N1 and N2, namely, it is the sum of the linear combination of Gaussian distribution by which weighting was carried out with the

predetermined weighting factor. Each weighting factor is set to ws11, ws12, wn1, and wn2. At this time, output distribution of the 1st condition of having been adaptation-ized after composition becomes the sum with a weighting factor of four Gaussian distribution, S11+N1, S12+N1, S11+N2, and S12+N2. Furthermore, HMM composition in the combination of the condition of a noise gauss mixing model and the 3rd condition of Voice HMM is similarly performed in the HMM composition in the combination of the condition of a noise gauss mixing model, and the 2nd condition of Voice HMM, and a list.

[0030] Therefore, the voice HMM in which HMM multi-pass composition section 13a does not have the noise stored in the voice HMM memory 32, The noise gauss mixing model stored in the model memory 22 is set in the condition of all the combination of each of these conditions using a well-known HMM synthesis method. It is adaptation-ized HMM including the mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor. Two or more adaptation-ized HMM(s) corresponding to two or more signal-to-noise ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / without the above-mentioned noise / above-mentioned] are generated. It outputs and stores in the HMM memory 23 which compounds and becomes so that adaptation-ized HMM of the generated plurality may be juxtaposed and it may become a multi-pass format and which generated adaptation-ized HMM and was adaptation-ized.

[0031] In drawing 1 , a voice recognition unit 200 is equipped with a microphone 1, A/D converter 2, the feature-extraction section 3, and the speech recognition section 4, and is constituted. After the generating voice of a natural utterance sentence is inputted into a microphone 1 and changed into an utterance sound signal, A/D conversion of it is carried out to a voice digital data signal with a predetermined sampling frequency by A/D converter 2. Subsequently, based on the voice digital data signal inputted, by carrying out LPC analysis, the feature-extraction section 3 extracts the feature vector containing the 12th mel cepstrum multiplier, 12th delta mel cepstrum multiplier, power, and delta power, and outputs it to the speech recognition section 4. Furthermore, while the speech recognition section 4 calculates the likelihood of a phoneme using adaptation-ized HMM which was stored in the HMM memory 23 The likelihood of a word is calculated using the word "HMM" of the predetermined phoneme base beforehand stored in the word HMM memory 5, by determining the word which consists of a phoneme from which output likelihood serves as max, speech recognition processing is performed, and the character string of the maximum ****** of a speech recognition result is generated and outputted.

[0032] <Operation gestalt of ** 2nd> drawing 2 is the block diagram showing the configuration of HMM model generation equipment 100b which is the 2nd operation gestalt concerning this invention, a voice recognition unit 200, and environmental noise measurement equipment 300, attaches the same sign about the same component as drawing 1 in drawing 2 , and omits those detail explanation. This 2nd operation gestalt has the following differences as compared with the 1st operation gestalt illustrated by drawing 1 .

(1) HMM model generation equipment 100b is replaced with HMM multi-pass composition section 13a as compared with HMM model generation equipment 100a, and it is equipped with the mixed weight adaptation-ized section 14 between the noise gauss mixing model memory 22 and the HMM composition section 13 while it is equipped with the HMM composition section 13 concerning the 3rd conventional example.

(2) It has further the ambient noise memory 310 connected with the ambient noise measuring device 300 at it. Hereafter, these differences are explained to a detail.

[0033] When using a voice recognition unit 200 under a real environment, it is not avoided that the ambient noise depending on a surrounding environment mixes in a microphone 1. Predicting is difficult for many of noises to mix, and the robust sound model is called for from mixing of the noise to change. With this operation gestalt, the HMM synthesis method incorporating adaptation-ization of the ambient noise model built with the noise database is used to this trouble. In the former, although study which used the real noise of an operating environment for generation of the model of an environmental sound was performed, since the amount of data of a real noise acquirable from practical constraint was restricted, the ambient noise model obtained from little data had the trouble of being weak, to fluctuation. So, with this operation gestalt, the initial ambient noise model is prepared using the noise database, and little real noise data perform adaptation-ization.

[0034] With this operation gestalt, the ambient noise at the time of speech recognition assumes it being added to generating voice by the imagination adder 6, and being inputted into a microphone 1. On the other hand, with this operation gestalt, by inputting ambient noise in case there is no generating voice into the ambient noise measuring device 300, and the ambient noise measuring device 300 concerned changing the ambient noise inputted into an electrical signal with a microphone, and carrying out A/D conversion with an A/D converter the digital data of ambient noise is obtained and it stores in the ambient noise memory 310. The digital data of this ambient noise is study data for adaptation-izing little in a short time which is 1 second.

[0035] The mixed weight adaptation-ized section 14 is based on the digital data of the ambient noise in which the mixed weight in the noise gauss mixing model of two or more mixing was stored by the ambient noise memory 310 in the one condition stored in the noise gauss mixing model memory 22, for example, is a well-known maximum-aposteriori-probability-estimation method (henceforth the MAP presuming method). For example, refer to the conventional technical reference 4 "Seiichi Nakagawa, "the speech recognition by the probability model", the Institute of Electronics, Information and Communication Engineers, pp.152-155, and July 1, Showa 63 first-edition issue". It uses, and adaptation-ization is performed and the adaptation-ized noise gauss mixing model is outputted to the HMM composition section 13 so that the maximum a-posteriori probability which is an example of likelihood may serve as max. Here, since adaptation-ization is limited to the weighting factor of a noise gauss mixing model, distribution or an average of each probability distribution do not change between the model which carried out HMM composition after performing adaptation-ization, and HMM compounded the first stage. Therefore, the adaptation-ization HMM can be obtained by matching with the initial composition HMM the weighting factor obtained by adaptation-ization. Subsequently, the HMM composition section 13 compounds the voice HMM without the noise stored in the voice HMM memory 32, and the noise gauss mixing model outputted from the mixed weight adaptation-ized section 14 using a well-known HMM synthesis method, and outputs and stores the HMM concerned compounded and adaptation-ized in the HMM memory 23 generated and adaptation-ized.

[0036] In the operation gestalt constituted as mentioned above, since mixed weight of a noise gauss mixing model is adaptation-ized based on the little ambient noise data at the time of speech recognition, the real noise amount of data which generation of an ambient noise model takes can be reduced sharply, and robustness over noise fluctuation can be made dogged at coincidence.

[0037] <Operation gestalt of ** 3rd> drawing 3 is the block diagram showing the configuration of HMM model generation equipment 100c which is the 3rd operation gestalt concerning this invention, a voice recognition unit 200, and environmental noise measurement equipment 300, attaches the same sign about the same component as drawing 1 and drawing 2 in drawing 3 , and omits those detail explanation. This 3rd operation gestalt has the difference of "having unified the mixed weight adaptation-ized section 14 and the HMM composition section 13, and having been referred to as formation of mixed weight adaptation, and HMM composition section 13b" as compared with the 2nd operation gestalt illustrated by drawing 2 . Hereafter, this difference is explained to a detail.
[0038] In this operation gestalt, after the formation of mixed weight adaptation and HMM composition section 13b carry out adaptation-ization like the 2nd operation gestalt, they perform processing of postadaptation-izing which carried out HMM composition rather than carry out HMM composition. That is, first, for next count simplification, the noise gauss mixing model stored in the noise gauss mixing model memory 22 and the voice HMM stored in the voice HMM memory 32 without a noise are compounded using an above-mentioned HMM synthesis method, and the initial composition HMM is prepared. Subsequently, HMM adaptation-ized by performing mixed weight adaptation-ization using the MAP presuming method is obtained to the initial composition HMM which made above-mentioned] preparations based on the digital data of little ambient noise in a short time stored in the ambient noise memory 310 on the occasion of environmental-adaptation-izing. Since adaptation-ization is limited to the mixed weighting factor of a noise gauss mixing model, it is only weighting factors the adaptation-ization HMM which carried out HMM composition after performing adaptation-ization, and for an average or distribution of each probability distribution not to change and to change with environmental adaptation-ization between the initial composition HMM. Therefore, the adaptation-ization HMM can be obtained by being directly reflected in the model after compounding the weighting factor obtained by adaptation-ization of a noise gauss mixing model, and computational complexity can be reduced greatly. This processing is shown in drawing 8 .
[0039] Although each component in adaptation-ized HMM carried out the multiplication of the mixed weight of a noise gauss mixing model to what compounded Voice HMM and a noise gauss mixing model in the spectral region, it is expressed with linear combination so that clearly from drawing 8 . Therefore, processing of postadaptation-izing which carried out HMM composition can be performed, and, thereby, computational complexity can be sharply reduced as compared with the 2nd operation gestalt.
[0040] <Operation gestalt of ** 4th> drawing 4 is the block diagram showing the configuration of 100d of HMM model generation equipment which is the 4th operation gestalt concerning this invention, a voice recognition unit 200, and environmental noise measurement equipment 300, attaches the same sign about the same component as drawing 1 thru/or drawing 3 in drawing 4 , and omits those detail explanation. This 4th operation gestalt has the difference of "having replaced with the formation of mixed weight adaptation, and HMM composition section 13b, and having had the formation of mixed weight adaptation, and HMM multi-pass composition section 13c" as compared with the 3rd operation gestalt illustrated by drawing 3 . Hereafter, this difference is explained to a detail.
[0041] The formation of mixed weight adaptation and HMM multi-pass composition section 13c concerning this operation gestalt are characterized by to use the HMM multi-pass synthesis method by HMM multi-pass composition section 13a concerning the 1st operation gestalt, when compounding a noise gauss mixing model and the voice HMM without a noise as compared with the formation of mixed weight adaptation and HMM composition section 13b concerning the 3rd operation gestalt.
[0042] Since according to the 4th operation gestalt constituted as mentioned above mixed weight of a noise gauss mixing model is adaptation-ized for a short time based on the digital data of little ambient noise and it is compounding using a HMM multi-pass synthesis method, it is adaptation-ized by ambient noise, and about an SN ratio, a degree of freedom is high and can generate the sound model which can respond to various SN ratios. Moreover, computational complexity is sharply reducible by performing "processing of postadaptation-izing which carried out HMM composition" concerning the 3rd operation gestalt.
[0043]
[Example] this invention persons use the HMM model generation equipments 100a, 100c, and 100d and voice recognition unit 200 concerning this operation gestalt. The sound model which is HMM which conducted the word recognition experiment of a continuation figure and was adaptation-ized by the engine performance AURORA2 database (for example) "The conventional technical reference 5 H.G.Hirsch et al and "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRWA SR2000, "Automatic Speech Recognition:Challenges for the Next Millennium", and refer to September, 2000." It used and evaluated. This AURORA2 database is a database for speech recognition system evaluations under a noise environment, and shows that detail in Table 1.
[0044]
[Table 1]
AURORA2 database ---------- task: -- continuation figure recognition sampling frequency [of 4 figures]: -- 8kHz16bitPCM / monophonic recording ---------- study set: -- noise: -- a subway and a crowd -- noisily -- noise, automobile noise, and show hole SN ratio: -- noise-less all 8840 [number of utterance:] 20dB 15dB 10dB 5dB ---------- Test set A:noise: -- a subway and a crowd -- noisily -- a noise, an automobile noise, show hole SN ratio:-5dB, and 0dB -- noise-less all 28028 [number of utterance:] 20dB 15dB 10dB 5dB ---------- Test set B:noise: -- the passage of a restaurant and shopping quarter -- station SN ratio:-5dB of an airport and a train, 0dB, 5dB, 10dB, 15dB, 20dB, and the noise-less all number of utterance -- :28028 ---------- [0045] the sound model (HMM without a noise) which created the result of the base line (1st example of a comparison) hereafter using the voice data which does not include a noise among study sets -- A sets of B sets of averages of all results are shown.
[0046] It is the speech recognition under a noise environment, and the method of building an easiest and ideal sound model is an approach of building a sound model using the study data under the same noise environment as input voice. The sound model hereafter created by the study data under the same noise environment as input voice is called "the model of the same noise" (2nd example of a comparison). The voice data which the noise which corresponded among A sets mixed is used for evaluation using the subset which one kind of noise mixed in study of the model of the same noise among study sets. Let these averages be the engine performance of the model of the same noise. It evaluates using the voice data which the noise which does not correspond to the sound model created as a model of the same above-mentioned noise as evaluation when the mixing noises of sound model study data and evaluation data differ among A sets mixed. A result is shown in drawing 9 . The WORD accuracy (%) which is an evaluation value here is defined by the

degree type.
[0047]
[Equation 3]
ワードアキュラシー（%）

$$= \frac{\text{全認識単語数} - (\text{置換誤り単語数} + \text{挿入誤り単語数} + \text{脱落誤り単語数})}{\text{全入力単語数}}$$

[0048] When the noise environment of study data and input voice is not in agreement so that clearly from drawing 9 , the recognition engine performance is falling sharply with the fall of the SN ratio of input voice.
[0049] The recognition experiment which used B sets of AURORA2 for evaluation data is conducted as evaluation of the formation of sound model environmental adaptation by the HMM synthesis method concerning the conventional technical reference 1 and 2. A noise model is learned using noise data (10 seconds) to each noise of evaluation data. Here, a noise model uses GMM of 1 condition 8 mixing. The following two sound models are created and compared using this noise model.
(1) SN ratio = the sound model which carried out HMM composition as 15dB (based on the HMM synthesis method concerning the conventional technical reference 1 and 2)
(2) SN ratio = the multi-pass-ized sound model which carried out HMM composition as 5, 10, 15, 20, and infinity(with no noise) dB (based on the HMM multi-pass synthesis method concerning the 1st operation gestalt)
[0050] These experimental results are also shown in drawing 9 . The result of the model of the same noise as the base line is also united and shown for a comparison. In SN ratio =15dB immobilization, by adaptation-ization by HMM composition, 13% of improvement in the engine performance was found. Moreover, in SN ratio =5dB, the engine performance high 58% was obtained by using multi-pass-ization of HMM compared with the base-line model.
[0051] Subsequently, the recognition experiment which used B sets of AURORA2 database for evaluation data is conducted as evaluation of the formation of sound model adaptation by the HMM composition concerning the 3rd and 4th operation gestalten. A noise model is set to GMM of 1 condition 8 mixing, and learns an initial noise model using the noise data for time amount length 10 second x25 kind, and sum total time amount length 250 seconds from the Japan Electronic Industry Development Association noise database (for example, conventional technical reference 3 reference.). Adaptation-ization of a noise model is performed to each noise of evaluation data using noise data (time amount length 1 second). The following sound models were generated using this noise model.
(1) SN ratio = the sound model which carried out HMM composition as 15dB (based on the HMM synthesis method concerning the 3rd operation gestalt)
(2) SN ratio = the multi-pass-ized sound model which carried out HMM composition as 5, 10, 15, 20, and infinity(with no noise) dB (based on the HMM multi-pass synthesis method concerning the 4th operation gestalt)
[0052] The experimental result to these sound models is also shown in drawing 9 . Using the synthesis method concerning the 3rd or 4th operation gestalt can attain the recognition engine performance almost equivalent to a conventional method by the adaptation amount of data of 1/10 so that clearly from drawing 9 . Moreover, in SN ratio =15dB immobilization, 14% of improvement in the engine performance was found to the base-line model. Furthermore, in SN ratio =5dB, 53% of improvement in the engine performance was obtained by multi-pass-ization of an adaptation-ized model again compared with the base line.
[0053] Therefore, the equipment using the synthesis method concerning this operation gestalt has the following characteristic effectiveness.
(1) Since the gauss mixing model is generated based on the wave signal data of the ambient noise of two or more classes, adaptation-ized HMM which compounded this gauss mixing model and Voice HMM serves as a dogged model to mixing of a strange noise.
(2) By using the model of two or more mixing as a noise model, construction of an effective noise model can be performed to various noises, and the resistance over time fluctuation of a noise improves.
(3) When the mixed weighting factor of a noise model is adaptation-ized using ambient noise data, as compared with the conventional example, computational complexity can be mitigated sharply, and environmental adaptation-ization is attained also in a large-scale sound model at a high speed. Moreover, the rate of speech recognition can be improved sharply.
(4) According to the multi-pass model classified by SN ratio, it is not restrained by constraint that the SN ratio of input voice is known, but the high rate of speech recognition can be obtained as compared with the conventional example at the time of speech recognition.
[0054]
[Effect of the Invention] As explained in full detail above, according to the sound model generation equipment concerning invention of the 1st of this application From the wave signal data of the ambient noise of two or more classes for study, so that output likelihood may serve as max In the condition of all the combination of the noise gauss mixing model which generated the gauss mixing model of two or more mixing in the one condition, and was generated by the voice HMM without a predetermined noise, and the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. And the sound model which compounds and becomes so that two or more adaptation-ized HMM(s) corresponding to two or more signal-to-noise ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / without the above-mentioned noise / above-mentioned] may be generated, adaptation-ized HMM of the generated plurality may be juxtaposed and i may become a multi-pass format is generated. Therefore, it is not restrained by constraint that the SN ratio of input voice is known. Moreover, as compared with the 3rd conventional example, the high rate of speech recognition can be obtained at the time of speech recognition, without being dogged and increasing the computational complexity of a sound model to mixing of a strange noise.
[0055] Moreover, according to the sound model generation equipment concerning invention of the 2nd of this application, from the wave signal data of the ambient noise of two or more classes for study, so that output likelihood may serve as max In the condition of all the combination of the noise gauss mixing model which generated the gauss mixing model of two or more mixing in the one condition, and was generated by the voice HMM without a predetermined noise, and the above-mentioned generation means to each of these conditions The sound model which becomes in HMM adaptation-ized in the mixed weight of the above-mentioned noise gauss mixing model based

on the ambient noise data at the time of speech recognition with the predetermined weighting factor, including the mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out is generated. Therefore, since it is adaptation-ized based on ambient noise data, as compared with the 3rd conventional example, the high rate of speech recognition can be obtained at the time of speech recognition, and to mixing of a strange noise, it is dogged and computational complexity of a sound model is not increased.

[0056] Furthermore, according to the sound model generation equipment concerning invention of the 3rd of this application, from the wave signal data of the ambient noise of two or more classes for study, so that output likelihood may serve as max In the condition of all the combination of the noise gauss mixing model which generated the gauss mixing model of two or more mixing in the one condition, and was generated by the voice HMM without a predetermined noise, and the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. Based on the ambient noise data at the time of speech recognition, mixed weight of the above-mentioned noise gauss mixing model is adaptation-ized. And the sound model which compounds and becomes so that two or more adaptation-ized HMM(s) corresponding to two or more signal-to-noise ratios between the noise gauss mixing models to which generation was carried out [voice / HMM / without the above-mentioned noise / above-mentioned] may be generated, adaptation-ized HMM of the generated plurality may be juxtaposed and it may become a multi-pass format is generated. Therefore, it is not restrained by constraint that the SN ratio of input voice is known. Moreover, as compared with the 3rd conventional example, the high rate of speech recognition can be obtained at the time of speech recognition, without being dogged and increasing the computational complexity of a sound model to mixing of a strange noise.

[0057] Moreover, according to the voice recognition unit concerning invention of the 4th of this application, the characteristic quantity is extracted based on the utterance sound signal of a natural utterance sentence, based on the extracted characteristic quantity, speech recognition of the above-mentioned utterance sound signal is performed using adaptation-ized HMM by which composition was carried out [above-mentioned], and a speech recognition result is outputted. Therefore, to the sound signal which the strange noise mixed, as compared with the conventional example, speech recognition can be carried out at the high rate of speech recognition, and a stout voice recognition unit can be offered to the voice which the noise superimposed.

[Translation done.]

CLAIMS

[Claim(s)]

[Claim 1] Sound model generation equipment characterized by providing the following. A storage means to store the wave signal data of the ambient noise of two or more classes for study A generation means to generate the gauss mixing model of two or more mixing in the one condition from the wave signal data of the ambient noise of two or more classes for study stored in the above-mentioned storage means so that output likelihood may serve as max A voice hidden Markov model without a predetermined noise In the condition of all the combination of the noise gauss mixing model generated by the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. And two or more adaptation-ized hidden Markov models corresponding to two or more S/N between the noise gauss mixing models by which generation was carried out [above-mentioned] with the voice hidden Markov model without the above-mentioned noise are generated. A synthetic means to generate the sound model which compounds and becomes so that the adaptation-ized hidden Markov model of the generated plurality may be juxtaposed and it may become a multi-pass format

[Claim 2] Sound model generation equipment characterized by providing the following. A storage means to store the wave signal data of the ambient noise of two or more classes for study A generation means to generate the gauss mixing model of two or more mixing in the one condition from the wave signal data of the ambient noise of two or more classes for study stored in the above-mentioned storage means so that output likelihood may serve as max A voice hidden Markov model without a predetermined noise A synthetic means generate the sound model which consists of a noise gauss mixing model generated by the above-mentioned generation means in the hidden Markov model adaptation-ized in the mixed weight of the above-mentioned noise gauss mixing model based on the ambient-noise data at the time of speech recognition, including the mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor in the condition of all the combination of each of these conditions

[Claim 3] Sound model generation equipment characterized by providing the following. A storage means to store the wave signal data of the ambient noise of two or more classes for study A generation means to generate the gauss mixing model of two or more mixing in the one condition from the wave signal data of the ambient noise of two or more classes for study stored in the above-mentioned storage means so that output likelihood may serve as max A voice hidden Markov model without a predetermined noise In the condition of all the combination of the noise gauss mixing model generated by the above-mentioned generation means to each of these conditions The mixed Gaussian distribution of each condition expressed with the sum of the linear combination of each Gaussian distribution by which weighting was carried out with the predetermined weighting factor is included. Based on the ambient noise data at the time of speech recognition, mixed weight of the above-mentioned noise gauss mixing model is adaptation-ized. And two or more adaptation-ized hidden Markov models corresponding to two or more S/N between the noise gauss mixing models by which generation was carried out above-mentioned] with the voice hidden Markov model without the above-mentioned noise are generated. A synthetic means to generate the sound model which compounds and becomes so that the adaptation-ized hidden Markov model of the generated plurality may be juxtaposed and it may become a multi-pass format

[Claim 4] The voice recognition unit characterized by to have an extract means extract the characteristic quantity based on the utterance sound signal of a natural utterance sentence, and a speech-recognition means perform speech recognition of the above-mentioned utterance sound signal using the sound model generated by claim 1 thru/or the sound model generation equipment of one of any 3 publications based on the characteristic quantity by which the extract was carried out [above-mentioned], and output a speech recognition result.

[Translation done.]